Final Report on the Pilot Data Warehouse Project

October 2, 2002

Executive Summary

Background

The Pilot Data Warehouse (PDW) Project is a project that was initiated in July 2001 by the Data Quality Action Team and the Division of Information Resources (DIR) to test assumptions and gain knowledge about issues that needed to be understood and addressed in order for IHS to implement an enterprise-wide, national data warehouse environment for IHS, Tribal, and Urban healthcare systems. (The entire Data Warehouse [DW] Project is thoroughly reviewed in the "DW-1 User Requirements/ Conceptual Design" document, Version 3, dated September 9, 2002.) This is the final report on the PDW portion of the larger, Data Warehouse Project.

The PDW was designed to accept a large spectrum of data extending from FY 1997 approximately through the end of FY 2001, data that came from multiple sources:

- Encounter and registration data from sites using RPMS (9 sites - 3 from each of 3 Areas)
- Encounter and registration data from a site not using RPMS
- Data from the Contract Health Services (CHS) Fiscal Intermediary (FI)
- CHS 638 data
- Dental data

IHS worked closely with experts from IBM and the SAS Institute on this project.

Because the primary purpose of the PDW project was to test assumptions and gain knowledge, it was designed along several specific principles. We designed the PDW to gather and store as much granularity as possible and we performed minimal data cleansing and data transformation in loading the staging tables and between staging and DW tables. We did not unduplicate encounter records if they come from different sources. Finally, we stored all modified, deleted records, flagging the current snapshot rather than deleting data.

Accomplishments

The PDW was remarkably successful in that we were able to test the most important assumptions and gained critical knowledge on most of the issues that we must understand and address. The following is what we accomplished during this project:

- ❖ Designed, tested, and implemented new logical and physical models
- ❖ Designed, tested, and implemented extraction, transformation, and load (ETL) processes to take data from the PDW exports, load them first into staging tables and finally into target tables consistent with those models.
- ❖ Made several important changes in the initial PDW design including converting the initial curser-based ETL process to an insert-based one and partitioning the database, markedly improving the load times.
- ❖ Loaded the following data types
  - ➢ RPMS encounter data
  - ➢ Modified RPMS encounter data from one site

  - RPMS registration data
  - RPMS third party eligibility data
  - RPMS CPT data
  - CHS Fiscal Intermediary data
  - CHS 638 (non-FI) data
  - Dental data
- Extensively analyzed data from one non-RPMS site to better understand the issues relating to loading such data in DW1.
- Designed and implemented post load processing, e.g., processes to handle suspect SSNs and names.
- Developed initial metadata on these data types, including detailed mappings from source system to target table fields, and then loaded that information into DB2 Data Warehouse Manager.
- Demonstrated that we could produce accurate Workload and User Pop Reports from the PDW.
- Demonstrated that we could produce reports providing information on the quality of their data back to source systems including reports on field content and deviations from historical norms.
- Demonstrated that we could produce ORYX and GPRA-type clinical performance measures from the PDW.
- Identified the types of ORYX data transport reports we would like to implement in DW1.
- Implemented an "unduplication" method for truly duplicated or modified RPMS encounter and registration records using their unique encounter and registration record IDs, respectively.
- Began a detailed exploration of unduplication methods across varied sources (RPMS and non-RPMS sites, CHS FI, CHS 638, Dental)
- Implemented a method using Integrity to unduplicate patient records both within Areas (NPIRS "regions") and nationally
- Identified updating of RPMS registration records as in important issue to be dealt with during the design phase of DW1.
- Identified data that will be handled in a more normalized fashion during DW1, e.g., medications, labs, and clinical measurements.
- Designed and implemented a backup and recovery process for PDW.
- Assessed and implemented a security plan for PDW.
- Produced extensive written documentation on all of the above

Based on our experience from this PDW project, the PDW team recommends to DIR senior management that we proceed with completing the design and implementation of the first iteration of an enterprise-wide, national data warehouse environment for IHS, Tribal, and Urban healthcare systems, as described in the "DW-1 User Requirements/ Conceptual Design" document.